PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (see an example) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below. Some articles will have been accepted based in part or entirely on reviews undertaken for other BMJ Group journals. These will be reproduced where possible.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Experiences with Global Trigger Tool reviews in five Danish hospitals – an implementation study
AUTHORS	von Plessen, Christian ; Kodal, Anne; Anhøj, Jacob

VERSION 1 - REVIEW

REVIEWER	Henriette Lipczak MD, Program Manager Danish Cancer Society, Quality & Patient Safety Strandboulevarden 49 DK - 2100 Copenhagen Denmark Competing interests: none
REVIEW RETURNED	25-Jun-2012

THE STUDY	Research question: the research question/aim is clearly stated in the abstract, but should be clarified further in the background section Study design: the study is mainly descriptive. A more structured approach towards data collection would have enhanced the quality – the anectdotal information could advantageously have been replaced by more systematic data collection using e.g. interview or questionnaire. Description of methods: The methods section commences with a description of the GTT reviews, however, a precise description of the study described in the paper would have been preferred, e.g. there is lack of information on the authors and their role in the study – who are they and how did they collect the information? This is key since the study is observational. Data on collection of background information is provided, but how was data on experiences with the implementation collected, by whom and when? Experienced nurses carried out the GTT review. Did they have experience in reviewing charts or did they have certain clinical experiences? A description of the consensus processes involving the two primary reviewers as well as the whole team (primary reviewers and physician) would contribute to the understanding of differences between the compared review teams, e.g. did the primary reviewers reach consensus in a face to face discussion, and did the three of them meet for the final decisions?

	The nurses who categorized harms – what clinical experience did they have? Were they affiliated with the hospitals where the harms were identified? The use of DPSD data in this study is not clear. The aim is to describe experiences with GTT – how does event reporting contribute to this? If used, it is not clear what period the DPSD data originates from.
RESULTS & CONCLUSIONS	Discussion: A discussion of the strengths and limitations of the study is missing. Unanswered questions and specific proposals for future research are not described – what elements of the GTT's measurement properties need further investigation?
REPORTING & ETHICS	The study is – in spite of its scientific limitations – highly relevant in that it provides insights of the variation of harm rates related to the use of GTT and calls for further research of a widely implemented tool that is used locally, nationally and internationally.
GENERAL COMMENTS	IHI state that in determining whether an adverse event occured, it should be considered that an AE is defined as unintended harm to a patient from the viewpoint to the patient. How was this managed? Did it lead to any discussions between nurses and physicians that supports the conclusion that nurses are more inclusive. It is stated that each AE/harm was assigned to only one type category. Does this indicate that sometimes a decision had to be made as to which of two or more categories to choose? If yes, what is the implication of this? It is stated in the discussion that the hospital with the highest PSI rate observed a significant increase in harms. This is interpreted as a result of a change in the culture of reporting and documentation. Are there any other possible explanations? Has the potential change in reviewer's attention once they have learned that specific event types occur frequently (e.g. pressure ulcers and gastrointestinal complications) been investigated? The differences in documentations systems are described, though not conferred any special significance. Is it possible that the layout/the presentation of data in the different systems affects the findings - that some data is more eyecatching in some systems than in another – or are layout properties the same across systems? If DPSD data are to be retained in the paper, a discussion of differences in chart review and event reporting would be of value; one would not expect the same types of events to be identified with the two methods – results must be interpreted in light of this. The discussion mainly concerns the variation in harm rates. It would also be interesting to learn more about the differences in the assessment of harm types and consequences, e.g. what would be the practical implications of these variations? And as a curiosum: how does a hospital use the knowledge that most harms are in the 'others' category for safety improvement? Although not stated directly in the conclusion, it is indicated that the authors do not interpret the results in a way that affect the decision to use the GTT. The authors recommend that health care staff and policy makers should be aware of the 'variation-problem'/the need for sufficient training and retraing of review teams – is it possible to

	specify this awareness; given the results of this study - what can GTT be used for – what shouldn't it be used for? What would be the 'dangers' if the limitations are not taken into account in the use of the GTT results? Also, is training and retraining of teams sufficient to consider GTT a valid tool and to recommend further use, even before further scientific evaluation of the measurement properties of the tool? A discussion of the idea of a 'global' measure of safety would be interesting – can a tool be 'global' if it does not measure omissions and is based on what is registered in the medical record (i.e does not include documentation errors and administrative processes leading to harm and don't take into account the patients' experience of patient safety)? Is it possible at all to develop a true global measure? Do any alternatives to using GTT exist and does electronic capture of triggers have the potential to reduce the problems described in this study?

REVIEWER	Ellen Tveter Deilkås MD PhD Clinical consultant/ Senior advisor, Akershus University Hospital/The Norwegian Knowledge Centre I have no competing interests.
REVIEW RETURNED	26-Jun-2012

THE STUDY	1. A) The aim of the study is to describe experiences with the Global Trigger Tool (GTT) in Denmark in order to identify ways to improve the performance of the GTT review teams. Since the study only presents characteristics and procedures of GTT teams at five hospitals and their quantitative results, and some anecdotal information, the objective of the study seems more to be to "present experiences with...." rather than to "describe "them". I suggest that the objective is adjusted accordingly. B) The authors argue that it is necessary to calibrate the GTT instrument and GTT teams before the instrument is adopted to evaluate safety performance in hospitals across health systems. It does not mention that the GTT manual (page 29) warns against using the instrument to compare results between hospitals: "The IHI Global Trigger Tool is meant to be used as a mechanism to track your organization's progress over time. Although efforts are made to maintain a standard of training and process for the IHI Global Trigger Tool, organizations will vary in the skill of reviewers and other aspects of the IHI Global Trigger Tool process. We assume this bias is relatively stable over time in a given organization. The stability over time allows comparison to your own organization over time, but is not as useful in comparing between organizations. You can use national data to determine if your rates are in the general range of others. Organizations that have decreased adverse event rates should also be contacted to learn how this was achieved, even if the data is not exactly the same as yours." To prevent the article from evaluating the instrument for a purpose it is not made to fulfill, it should be precise about what the purpose of the instrument is. 2. A)The conclusion draws support from a recent study from Sweden

	which has studied interrater reliability between five teams from different hospitals. Harm rates between these teams ranged from 27,2 to 99,7 per 1000 patient days, with a pairwise interrater reliability ranging from a kappa value of 0,26 to 0,77. The article does not mention that the team in the Swedish study, with the highest harm rate, team IV, used a different definition for harm (The Swedish National Board of Health and Welfare's definition of AE: 'Any suffering, discomfort, bodily or mental injury, illness or death caused by healthcare and which is not an inevitable consequence of the patient's condition or an expected effect of the treatment received by the patient because of her/his condition'), than the other four teams, which used the GTT definition of harm. The harm rates between the four teams which did use the same definition for harm, ranged from 27,2 to 33,2 AE's per 1000 patient days, with a pairwise interrater reliability kappa value estimate of 0,62 ranging from 0,38 to 0,81. These results are not that bad, and should be taken into account in the discussion. B)The article with reference number 19, referred to on page 12, is imprecisely cited (variation of harm by hospital was between 23, 1% and 37, 9%, and not 19,4 5 and 37,9%).
RESULTS & CONCLUSIONS	1 A) The results in the study are presented with only one of the measures that the instrument provides: Patient safety incidence per 1000 patient days. Since the intention of the article is to present experiences with the Global Trigger Tool, it should also present the results for the two other measurements, which the GTT provides: Adverse events per 100 admissions; and Percent of admissions with an adverse event. B) The study concludes that the way the GTT teams perform the reviews, strongly contribute to the differences in harm rates between the hospitals and suggests measures to improve and standardize the conditions for the GTT teams. It should be clear about that the differences in results between hospitals, and the reasons for them, do not contradict the purpose of the instrument.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

Research question:
The research question/aim is clearly stated in the abstract, but should be clarified further in the background section
###
We have rephrased the last paragraph of the introduction to clarify this point.
###

Study design:
The study is mainly descriptive. A more structured approach towards data collection would have enhanced the quality – the anecdotal information could advantageously have been replaced by more systematic data collection using e.g. interview or questionnaire.
###
We agree, a more systematic collection of contextual/anecdotal data, e.g. with a questionnaire, would have yielded more reliable results. However at this stage, we found an explorative approach adequate to collect contextual information on the training and review processes of the GTT teams. We have added some detail on how we collected the data in the method section to clarify the point.

###

Description of methods:
The methods section commences with a description of the GTT reviews, however, a precise description of the study described in the paper would have been preferred, e.g. there is lack of information on the authors and their role in the study – who are they and how did they collect the information? This is key since the study is observational.
###
Thank you for this comment. We have rearranged the paragraphs in the methods section and give more detailed information regarding these relevant questions.
###

Data on collection of background information is provided, but how was data on experiences with the implementation collected, by whom and when?
###
We have added this information in the methods section.
###

Experienced nurses carried out the GTT review. Did they have experience in reviewing charts or did they have certain clinical experiences?
###
The authors know most of the nurses doing the reviews but we did not systematically collect specific information on how experienced they were and have thus deleted the word “experienced” from the paragraph.
###

A description of the consensus processes involving the two primary reviewers as well as the whole team (primary reviewers and physician) would contribute to the understanding of differences between the compared review teams, e.g. did the primary reviewers reach consensus in a face to face discussion, and did the three of them meet for the final decisions?
###
There were differences in how the teams reviewed met. We have expanded the description of the process in the text otherwise the data are presented in Table 2. Further detail regarding this question would require new interviews.
###

The nurses who categorized harms – what clinical experience did they have? Were they affiliated with the hospitals where the harms were identified?
###
The two nurses are experienced clinicians; they work as quality coordinators at two departments of one of the hospitals participating in the Safer Hospital program and have extensive experience with GTT reviews. We have added the latter information in the methods section of the manuscript.
###

The use of DPSD data in this study is not clear. The aim is to describe experiences with GTT – how does event reporting contribute to this? If used, it is not clear what period the DPSD data originates from.
###
We present the PSI reporting rates as background information that should be considered a measure of safety culture at the hospitals. We have moved this part to the background in the methods section, added the time period (2010) and an explanation to make this point clearer.
###

Discussion:
A discussion of the strengths and limitations of the study is missing. Unanswered questions and specific proposals for future research are not described – what elements of the GTT's measurement properties need further investigation?
###
We agree, and we have added paragraphs with strengths/limitations and future research.
###

The study is – in spite of its scientific limitations – highly relevant in that it provides insights of the variation of harm rates related to the use of GTT and calls for further research of a widely implemented tool that is used locally, nationally and internationally.
###
9
###

IHI state that in determining whether an adverse event occurred, it should be considered that an AE is defined as unintended harm to a patient from the viewpoint to the patient. How was this managed? Did it lead to any discussions between nurses and physicians that support the conclusion that nurses are more inclusive.
###
An interesting question, but unfortunately we do not have this information.
###

It is stated that each AE/harm was assigned to only one type category. Does this indicate that sometimes a decision had to be made as to which of two or more categories to choose? If yes, what is the implication of this?
###
Yes, the two nurses discussed these cases and usually came to an agreement. When they could not the case was discussed with one of the authors, CvP. So, an element of judgment is inherent to this approach. On the other hand the same nurses categorized all harms from all hospitals. Moreover we used categories of harms that make sense clinically and are documented in the literature.
###

It is stated in the discussion that the hospital with the highest PSI rate observed a significant increase in harms. This is interpreted as a result of a change in the culture of reporting and documentation. Are there any other possible explanations? Has the potential change in reviewer's attention once they have learned that specific event types occur frequently (e.g. pressure ulcers and gastrointestinal complications) been investigated?
###
One would assume that reviewers can become biased towards frequent or otherwise "prominent" harms. On the other hand, the use of triggers should at least partly prevent such a development. This would be interesting idea for further research. We have changed the paragraph slightly to make it clearer.
###

The differences in documentations systems are described, though not conferred any special significance. Is it possible that the layout/the presentation of data in the different systems affects the findings - that some data is more eyecatching in some systems than in another – or are layout properties the same across systems?
###
We have not studied this aspect of the reviews but it would be an interesting question for further

research. The point has been added in the discussion section of the manuscript.
###

If DPSD data are to be retained in the paper, a discussion of differences in chart review and event reporting would be of value; one would not expect the same types of events to be identified with the two methods – results must be interpreted in light of this.
###
As mentioned above, we use the DPSD data as background and did not intend to compare the two approaches.
###

The discussion mainly concerns the variation in harm rates. It would also be interesting to learn more about the differences in the assessment of harm types and consequences, e.g. what would be the practical implications of these variations? And as a curiosum: how does a hospital use the knowledge that most harms are in the 'others' category for safety improvement?
###
We agree, that it would be very interesting to study the use of harm types or, as some of the hospitals in the Safer Hospital Program call it, the harm profiles but we do not have data to explore this question. As you point out the 'others' category is not useful for preventing adverse events. However, for the purpose of this manuscript we found it necessary to limit the number of categories.
###

Although not stated directly in the conclusion, it is indicated that the authors do not interpret the results in a way that affect the decision to use the GTT. The authors recommend that health care staff and policy makers should be aware of the 'variation-problem'/the need for sufficient training and retraining of review teams – is it possible to specify this awareness; given the results of this study - what can GTT be used for – what shouldn't it be used for? What would be the 'dangers' if the limitations are not taken into account in the use of the GTT results? Also, is training and retraining of teams sufficient to consider GTT a valid tool and to recommend further use, even before further scientific evaluation of the measurement properties of the tool?
###
In our opinion the study indicates caveats that should be considered when implementing the GTT in a new setting rather than limitations of the method as such. Therefore a discussion about the properties of the GTT as a method was not included in the paper. However, our findings support the recommendation that the GTT should not be used to compare hospitals. We have added a sentence on this in the concluding paragraph.
###

A discussion of the idea of a 'global' measure of safety would be interesting – can a tool be 'global' if it does not measure omissions and is based on what is registered in the medical record (i.e does not include documentation errors and administrative processes leading to harm and don't take into account the patients' experience of patient safety)? Is it possible at all to develop a true global measure?
###
An interesting discussion, but the aim of our paper was to describe the experiences with the GTT method as it is. We feel that a critique of the method as such is beyond the scope of the paper.
###

Do any alternatives to using GTT exist and does electronic capture of triggers have the potential to reduce the problems described in this study?
###

We find that alternative methods to identify harms and electronic trigger capture, although important and relevant areas, reach beyond the scope of this paper which aims to support teams in implementing the manual GTT reviews.
###

2. Reviewer: Ellen Tveter Deilkås MD PhD
Clinical consultant/ Senior advisor,
Akershus University Hospital/The Norwegian Knowledge Centre

I have no competing interests

1. A) The aim of the study is to describe experiences with the Global Trigger Tool (GTT) in Denmark in order to identify ways to improve the performance of the GTT review teams. Since the study only presents characteristics and procedures of GTT teams at five hospitals and their quantitative results, and some anecdotal information, the objective of the study seems more to be to "present experiences with...." rather than to “describe “them”. I suggest that the objective is adjusted accordingly.
###
We agree, and have changed the wording of the aim of the study.
###

B) The authors argue that it is necessary to calibrate the GTT instrument and GTT teams before the instrument is adopted to evaluate safety performance in hospitals across health systems. It does not mention that the GTT manual (page 29) warns against using the instrument to compare results between hospitals: “The IHI Global Trigger Tool is meant to be used as a mechanism to track your organization’s progress over time. Although efforts are made to maintain a standard of training and process for the IHI Global Trigger Tool, organizations will vary in the skill of reviewers and other aspects of the IHI Global Trigger Tool process. We assume this bias is relatively stable over time in a given organization. The stability over time allows comparison to your own organization over time, but is not as useful in comparing between organizations. You can use national data to determine if your rates are in the general range of others. Organizations that have decreased adverse event rates should also be contacted to learn how this was achieved, even if the data is not exactly the same as yours.”
To prevent the article from evaluating the instrument for a purpose it is not made to fulfill, it should be precise about what the purpose of the instrument is.
###
We agree that GTT is not a benchmarking tool. However, we were truly surprised how large the differences in harm rates in the five hospitals were. We have changed the discussion section to clarify this aspect.
###

2. A) The conclusion draws support from a recent study from Sweden which has studied interrater reliability between five teams from different hospitals. Harm rates between these teams ranged from 27,2 to 99,7 per 1000 patient days, with a pairwise interrater reliability ranging from a kappa value of 0,26 to 0,77. The article does not mention that the team in the Swedish study, with the highest harm rate, team IV, used a different definition for harm (The Swedish National Board of Health and Welfare’s definition of AE: ‘Any suffering, discomfort, bodily or mental injury, illness or death caused by healthcare and which is not an inevitable consequence of the patient’s condition or an expected effect of the treatment received by the patient because of her/his condition’), than the other four teams, which used the GTT definition of harm. The harm rates between the four teams which did use the same definition for harm, ranged from 27,2 to 33,2 AE’s per 1000 patient days, with a pairwise

interrater reliability kappa value estimate of 0,62 ranging from 0,38 to 0,81. These results are not that bad, and should be taken into account in the discussion.
###
Thank you for pointing this out. As you mention, the Swedish definition of an AE is somewhat broader than the GTT definition of harm which could explain some of the difference between team IV and the other teams. Notably, the team with the highest harm rate had not attended the network meetings with the other teams. The reference has been omitted from the manuscript.
###

B) The article with reference number 19, referred to on page 12, is imprecisely cited (variation of harm by hospital was between 23, 1% and 37, 9%, and not 19,4 5 and 37,9%).
###
Thank you, we mistook the lower CI for the value.
###

1 A) The results in the study are presented with only one of the measures that the instrument provides: Patient safety incidence per 1000 patient days. Since the intention of the article is to present experiences with the Global Trigger Tool, it should also present the results for the two other measurements, which the GTT provides: Adverse events per 100 admissions; and Percent of admissions with an adverse event.
###
We used the most widely known measure of harms per 1000 bed days because the focus of the paper was on the variation of harms across hospitals. We have added the harms per 100 admissions for each hospital under results. In our view the percentage of harmed patients, while useful in quality improvement, does not add much information with regard to the topic of this article. However, we can calculate these rates but will need more time than the editor permitted because of the summer holiday in Denmark.
###

B) The study concludes that the way the GTT teams perform the reviews, strongly contribute to the differences in harm rates between the hospitals and suggests measures to improve and standardize the conditions for the GTT teams. It should be clear about that the differences in results between hospitals, and the reasons for them, do not contradict the purpose of the instrument.
###
We agree and have changed parts of the discussion accordingly.
###

VERSION 2 – REVIEW

REVIEWER	Deilkås , Ellen Norwegian Knowledge Center, National Unit for Patient Safety I have no competing financial interests. I have designed, and trained the GTT teams that have performed, the national record review with GTT, in Norway. I may therefore be considered to have intellectual interests at stake related to this review. I leave to the editors to decide if that is the case.
REVIEW RETURNED	31-Jul-2012

THE STUDY	1.Page 29 line 28: It would strengthen the relevance of the paper if the third GTT measure, rate of harmed patients, also is mentioned, when describing the GTT method, since this has been the main measure in two national GTT reports from Norway and the US.

	2.Pge 39 line 34: Would the consequences of the methodological variability between the hospitals reported in this study, be reduced by using the GTT measure, rate of patients harmed? 3.Grammatical Page 40 line 24: A grammatical mistake makes the sentence unclear. 4.Grammatical correction Line 48 page 40: The strength is its relevance to the implementation… 5.Grammatical correction Line 50: Our contextual data are detailed and thus practical…
RESULTS & CONCLUSIONS	6. The GTT results are not complete with only one of the GTT measures presented. See above. 7.Page 40 line 18, Since the study has not statistically proven relations between diifferences in training, review processes, documentation routines and variations in rates of harm as measured by the GTT, it would be more appropriate if the authors state what they believe is a probable relation, rather than state a relation that has not been proven. I suggest that the first statement in the conclusion is modified accordingly. 8.Page 40 line 26: The word " finding" gives an association to a study that is based on statistical evidence rather than observational evidence. It would perhaps be more suitable to use the word observations, rather than the word findings.
GENERAL COMMENTS	I recommend that the authors are given the time necessary to supplement reported data with data from the GTT measure; rate of patients harmed.

VERSION 2 – AUTHOR RESPONSE

Replies to comments from reviewer

1. Page 29 line 28: It would strengthen the relevance of the paper if the third GTT measure, rate of harmed patients, also is mentioned, when describing the GTT method, since this has been the main measure in two national GTT reports from Norway and the US.
Reply:
We have collected the percentage of harmed patients from the databases of the five participating hospitals. We have gone through all the data again to make sure that we have the same primary source of data for all measures. This has led to minor changes in the harm rates that do not change the conclusions of the article. The shift at Hillerød hospital was not significant anymore and we have removed the sentence from the results section. We also looked over the classification of harms and the harm categories and have added a sentence on the documentation of the coding of the types of harm. All figures are revised and we have added a figure for the percentage of harmed patients.

2. Page 39 line 34: Would the consequences of the methodological variability between the hospitals reported in this study, be reduced by using the GTT measure, rate of patients harmed?
Reply
The percentage of harmed patients varied between 18 and 33% (1,7 fold), the harm rate between 34 and 84 per 1000 patient days (2,5 fold). As expected the variability is lower in the percentage of harmed patients.

3. Grammatical

Page 40 line 24: A grammatical mistake makes the sentence unclear.
Reply
Fragment of earlier sentence - deleted

4. Grammatical correction
Line 48 page 40: The strength is its relevance to the implementation…
Reply
Changed accordingly.

5. Grammatical correction
Line 50: Our contextual data are detailed and thus practical…
Reply
Changed accordingly.

6. The GTT results are not complete with only one of the GTT measures presented. See above.
Reply
Percentage of harmed patients added.

7. Page 40 line 18, Since the study has not statistically proven relations between differences in training, review processes, documentation routines and variations in rates of harm as measured by the GTT, it would be more appropriate if the authors state what they believe is a probable relation, rather than state a relation that has not been proven. I suggest that the first statement in the conclusion is modified accordingly.
Reply
We have added the word “probably” in the conclusion of the discussion and in the abstract.

8. Page 40 line 26: The word “finding” gives an association to a study that is based on statistical evidence rather than observational evidence. It would perhaps be more suitable to use the word observations, rather than the word findings.
Reply (Page 41, line 26
The wording of the sentence changed